# Implementing the data preservation and open access policy in CMS



2014 J. Phys.: Conf. Ser. 513 042029

(http://iopscience.iop.org/1742-6596/513/4/042029)

Journal of Physics: Conference Series **513** (2014) 042029 doi:10.1088/1742-6596/513/4/042029

# Implementing the data preservation and open access policy in CMS

**K Lassila-Perini[1], G Alverson[2], I Cabrillo[3], A Calderon[3], D Colling[4], M Hildreth[5,6], A Huffman[4], T Lampén[1], P Luukka[1], J Marco[3], T McCauley[5,6] and L Sonnenschein[7]**

[1] Helsinki Institute of Physics, Helsinki, Finland
[2] Northeastern Univ., Massachusetts, Boston, USA
[3] IFCA, CSIC-Univ. de Cantabria, Santander, Spain
[4] Department of Physics, Imperial College, London, UK
[5] University of Notre Dame, Notre Dame, IN, USA
[6] Fermi National Accelerator Laboratory, Batavia, IL, USA
[7] RWTH Aachen Univ., III. Physik. Inst. A,Aachen, Germany

E-mail: k.lassila-perini@cern.ch

**Abstract.** Implementation of the CMS policy on long-term data preservation, re-use and open access has started. Current practices in providing data additional to published papers and distributing simplified data-samples for outreach are promoted and consolidated. The first measures have been taken for analysis and data preservation for the internal use of the collaboration and for open access to part of the data. Two complementary approaches are followed. First, a virtual machine environment, which will pack all ingredients needed to compile and run a software release with which the legacy data was reconstructed. Second, a validation framework, maintaining the capability not only to read the old raw data, but also to reprocess them with an updated release or to another format to help ensure long-term reusability of the legacy data.

## 1. Introduction

CMS has approved a data preservation, re-use and open access policy[1, 2], which motivates and defines the CMS approach to preservation of the data and access to them at various levels of complexity. The implementation of the policy has been elevated to a dedicated project within the collaboration, covering areas from the analysis and data preservation for the internal use of the collaboration to open access to part of the data. CMS is looking for solutions, which could be usable for the other LHC experiments, and promotes common infrastructures wherever possible.

## 2. Publications and additional data

The CMS data policy emphasizes the importance of the possibility of re-use of the CMS data. In parallel to the open access papers, publication of additional numerical data in a form in which they can be easily extracted for further use is encouraged, as well as initiatives such as RIVET[3] allowing for easy comparison between observed data and Monte Carlo event generators and their validation. An example of a data table attached to a published paper is shown in figure 1. For

Table 1 ( P 7,10. ) HIDE DATA or as: plain text, AIDA, PyROOT, YODA, ROOT, mpl, ScaVis or MarcXML
The fiducial and acceptance-corrected cross sections for PT<50 GeV/c and |rapidity|<2.4.

Additional systematic error: ± 4.0% (luminosity uncertainty)

| | ABS(YRAP) : < 2.4 | |
|---|---|---|
| | PT : < 50 GeV | |
| | SQRT(S) : 7000.0 GeV | |
| | FIDUCIAL | ACCEPTANCE-CORRECTED |
| RE | SIG IN NB | |
| | | HIDE DATA |
| P P --> UPSI(1S) < MU+ MU-> X | 3.06 ± 0.20 (stat) +0.20,-0.18 (sys) | 8.55 ± 0.05 (stat) +0.56,-0.50 (sys) |
| P P --> UPSI(2S) < MU+ MU-> X | 0.910 ± 0.011 (stat) +0.055,-0.046 (sys) | 2.21 ± 0.03 (stat) +0.16,-0.14 (sys) |
| P P --> UPSI(3S) < MU+ MU-> X | 0.490 ± 0.010 (stat) ± 0.029 (sys) | 1.11 ± 0.02 (stat) +0.10,-0.08 (sys) |
| | Plot SelectPlot | Plot SelectPlot |

**Figure 1.** An example of additional data table attached to a publication.

the distribution of these data, CMS relies on digital library services such as CDS, INSPIRE and HEPData.

## 3. Data for outreach and education

From March 2010 the CMS collaboration had agreed to the release of several datasets. This release is for outreach and education and is in a simplified format. The agreed release is 300,000+ events including electrons, muons, J/psis, Upsilons, W and Z bosons, and Higgs candidate events. More information on the dataset content as well as access to the data itself can be found on the CMS website[4]. The most widely-used use-case has been in the masterclasses developed and organized by QuarkNet[5] and the International Particle Physics Outreach Group (IPPOG)[6].

Various tools are available for examination of the events. An online event display developed by I2U2 is available[7]. This display is based on the look-and-feel and functionality of the iSpy event display[8] and reads the ig file format: each event is in a text-based JSON format and an ig file is a .zip archive. The ig files are created using the software framework of the CMS experiment, converting the CMS format into ig format. This online display software is written in JavaScript and only requires a modern browser for use. An example display of a Higgs candidate event is shown in Figure 2. Other online client-side tools can be developed from csv and JSON summary files generated from the ig files.

## 4. Legacy data sets

Data are archived at many levels. At the raw data level, two distinct copies always exist and the custodial responsibility is within T0 at CERN and the T1 computing site storing the data. No raw data will be deleted. At the reconstructed data level, the data are stored in the Analysis Object Data (AOD) format, which is a subset of the reconstructed (RECO) data objects in an event. AOD alone is sufficient for most kinds of physics analysis and is the data format in which the CMS experiment stores the legacy data for long-term preservation purposes. All 2011-2012

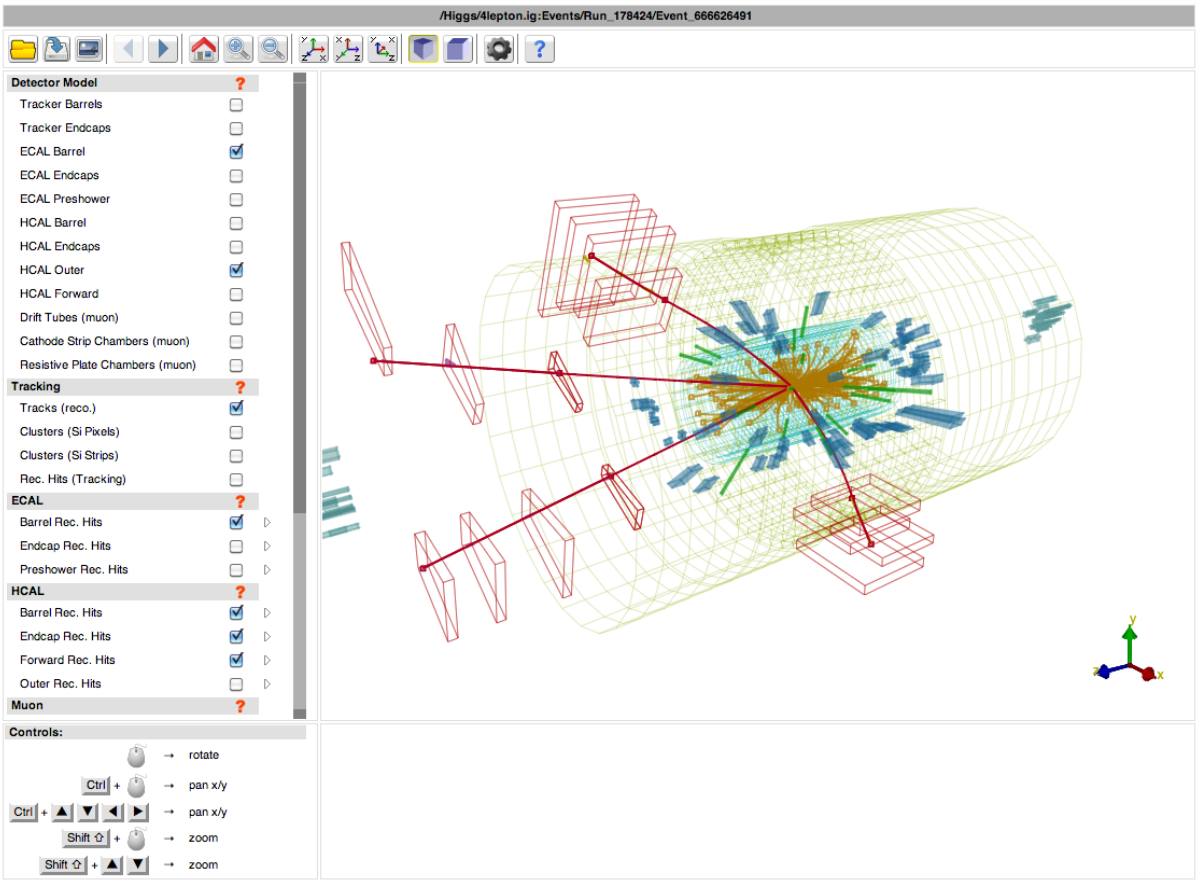


**Figure 2.** A Higgs candidate event shown with the online event display.

collision and simulated data are being reprocessed to a legacy data set in AOD format with a single CMS Software version.

Duplicates of reprocessed data sets are regularly deprecated, and when all active analyses (and eventual work for e.g. thesis preparation) have finished, the reprocessed data sets preceding the final legacy data set can be removed, after a consultation within the physics analysis groups of CMS.

## 5. Bit-level data preservation

At the bit-level, while the CMS computing model offers a solid base for long-term data preservation, CMS is looking forward to the program of work of the bit-level data preservation working group under HEPiX, where the challenges in this area will be discussed in the context of the Worldwide LHC Computing Grid (WLCG). It will cover aspects such as technology used for long-term archiving, definition of reliability, mitigation of data loss risks, monitoring and verification of the archive contents, procedures for recovering unavailable and/or lost data, and procedures for archive migration to new-generation technology, and the working group will provide recommendations for sustainable archival storage of HEP data across multiple sites and different technologies.

## 6. Analysis preservation

At the level of reconstructed data and its re-use, the experience from the other HEP experiments indicates that the biggest challenge in data preservation is the loss of knowledge and expertise. While CMS, and HEP experiments in general, do very well in recording "immediate" metadata, such as event and run numbers, beam conditions, software versions used in the data reprocessing, we are doing poorly on "context" metadata, i.e. practical information needed to put the data in context and analyze them. Therefore, CMS is actively looking for solutions to enable the easy recording of this detailed knowledge, readily available at the time of the active physics analysis, but quickly forgotten. The Invenio team at CERN, with input from CMS, will setup a prototype of a tool which could make recording and documenting of the relevant details easy. This documentation can be restricted to the collaboration members only if required, and only the parts considered appropriate could be available in the public domain.

CMS is also active in Data and Software Preservation for Open Science, DASPOS[9], which represents an initial exploration of the key technical problems that must be solved to provide appropriate data, software and algorithmic preservation for HEP, including the contexts necessary to understand, trust and reuse the data. While the archiving of HEP data may require some HEP-specific technical solutions, DASPOS will create a template for preservation that will be useful across many different disciplines, leading to a broad, coordinated effort.

## 7. Long-term validation

A framework to enable long-term validation of CMS data is being prepared. The goal of the long-term validation project is to extract and record the necessary information and tools in order to be able to validate the data and software in case of re-use in long-term future. The powerful tools already in use in CMS for validation of software releases and for data quality monitoring will be used in the long-term validation project. A set of reference plots covering physics results is being defined: starting from individual objects (muons, electrons,...) to high level physics signatures involving basic analysis selections. The first exercise is proceeding based on 2010 data, a part of which will be publicly released.

## 8. Open access

CMS is currently preparing for a public release of part of the 2010 collision and simulated data in AOD format. We expect that regular public releases will be made of the data after they have been analyzed by the collaboration. The 2010 data release will be based on the latest reprocessing of the full data set, which took place in spring 2011. The future releases will be based on well-defined legacy data sets, which are at the moment being reprocessed for the 2011-2012 data.

Access to the data requires a compatible computing environment, which in the case of the first public release will be based on software versions running on the SLC5 operating system. To make things easier for open access users, a virtual machine (VM) image has been prepared, in a format usable by the freely-available VirtualBox[10] application. For access to the data, the initial work flow is kept as close to the standard one as possible, which uses xroot. An xrootd[11] server has been commissioned with anonymous read-only access, further limited by firewall to include only those sites involved in the testing phase.

The AOD files and format are appropriate for physics analysis, and the public release will be accompanied by stable, open source software needed for a number of example analysis and suitable documentation. However, the AOD files and the analysis software are intrinsically complex and good knowledge is needed in order to perform a full physics analysis including proper estimates of the uncertainties.

Therefore, to identify the tools and instructions necessary to bring the data to a wide audience, preparing high-school level classroom applications has been chosen as a pilot use-

case[12] in the context of a larger project of the Finnish Ministry of Education to bring research data into open use[13]. For applications at the high-school level, an intermediate step is required in which the AOD format is processed into a format appropriate for the classroom user environment. Tools for this intermediate processing, including tools for validation of the newly processed format, will be provided, and it is expected that these tools, requiring setup of a CMS software environment, will be used by national and local data centres, who will then distribute the reprocessed data for end-users. The novelty of this approach compared to the release of small pre-selected data samples is the full open access, i.e. the possibility for non-members of the CMS collaboration to have direct access to the original research data and to the tools to further process and analyze them.

As the time and manpower constraints limit the possibilities of reprocessing the full 2010 data sets with the same legacy version of the 2011-2012 data, it is likely that the VM environment prepared for open access will serve as the entry point to the 2010 data also for the CMS collaborators. The efforts put in documenting the data and in setting up a concise set of instructions for their use will benefit end-users both within and outside of the collaboration.

## 9. Outlook

Data preservation must start when the data are created - CMS aims to achieve this goal by defining clear procedures and adopting easy-to-use tools, with which data preservation will be made possible without a heavy additional workload. CMS works in collaboration with the DPHEP (Data Preservation in High Energy Physics) collaboration, the CERN-IT department and other laboratories to define the tools and services covering the areas and levels of data preservation mentioned above. CMS seeks for generic solutions, making possible the use of common approaches with other experiments and disciplines.

## References

[1] CMS data preservation, re-use and open access policy: https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=6032
[2] Preparing for long-term data preservation and access in CMS, K Lassila-Perini et al 2012 *J. Phys.: Conf. Ser.* 396 032065
[3] https://rivet.hepforge.org/
[4] http://cms.web.cern.ch/content/cms-public-data
[5] http://quarknet.fnal.gov/
[6] http://ippog.web.cern.ch/
[7] A browser-based event display for the CMS experiment at the LHC, M Hategan et al 2012 *J. Phys.: Conf. Ser.* 396 022022; ; http://www.i2u2.org/elab/cms/event-display
[8] iSpy: a powerful and lightweight event display, G Alverson et al 2012 *J. Phys.: Conf. Ser.* 396 022002
[9] https://daspos.crc.nd.edu/
[10] https://www.virtualbox.org/
[11] http://xrootd.org/index.html
[12] Open CMS Data Finland: https://twiki.cern.ch/twiki/bin/view/HIPCMSExperiment/CMSOpenDataProject
[13] http://www.csc.fi/sivut/tta/national-reseach-data-project